ExonTracker™

Version 2.0

User's Manual

© 2003 GeneHarbor Inc. www.geneharbor.com

Copyrights

© 2003 GeneHarbor Inc. All rights reserved.

GeneHarbor, ExonTracker™ and ".xon" are the trademarks of GeneHarbor Inc.

All other trademarks and registered trademarks are property of their respective owners.

License Agreement

GeneHarbor, Inc. grants a license to use the accompanying software and printed material to you, the original purchaser. This is a binding Agreement between you and GeneHarbor Inc.. Use of the software shall constitute your acceptance of this Agreement. The copying of the software is strictly prohibited and adherence to this requirement is your sole responsibility. GeneHarbor Inc. reserves the right to modify and update the software and printed material without obligation to notify you, the original owner, of any change in the software and printed material.

Limited Warranty

If the performance of the software does not meet the standard described in the documentation, GeneHarbor Inc. will replace the software if notified within 30 days of purchase. In the event of a replacement agreed by GeneHarbor Inc., the original software, accessories, and documentation must be received by GeneHarbor Inc. in order for a replacement to be sent to you. The original users can replace a new version of software free of charge if there is one available. Under no circumstances shall GeneHarbor Inc. and its officers or its distributors be liable for any indirect, incidental, consequential or exemplary damages arising from the use, or the inability to use the software even if they were aware of the possibility of such damages.

GeneHarbor Inc.

info@geneharbor.com

www.geneharbor.com

Table Contents

Copyrights & License Agreement……... 1
Introduction…… 3
CD-ROM Installation…… 4
Overview……. 5

Protocol

I. Preparation for launching the genomic BLAST……..… 7
II. Genomic BLAST……... 10
III. Data Processing and Integration…….…… 11
IV. Addition of Protein Domain Information ……..…… 14
V. Data Export……...……..….. 17
VI. Non-Similarity Search…… 20
VII. Primer Design – Take Exon Junctions Into Consideration……. 23

Appendix

A. Web Link Update…….. 26
B. Manual Selection of A Coding Region …… 27
C. Set Exon and Intron Scales…… 28

Introduction

Several years ago, the human genome project was completed. The direct outcome of the project is that the arrangement of the base pairs of nucleotides comprising human genome becomes completely known. Its historic significance may not be understood fully at present time, but it immediately provides an answer to a very important question, that is, how many genes exist in human genome? To many of us a big surprise, the results coming from the study by several groups indicate that we human have only about 35,000 to 45,000 genes, much fewer than what people originally thought. This number is not much significant higher than that from a lower eukaryotic organism, such as Fly.

Considering the much complex human body, the amount of genetic information required for human should be conceivably much larger than that for any lower eukaryotic organism. The answer to the discrepancy in the genetic information may very well be due to the ways of the regulations of gene expressions. One of the major differences between eukaryotic organisms and prokaryotic organisms is that the majority of mRNAs in eukaryotic organisms are transcribed as pre-RNAs from intron-containing genes and processed into mature RNAs through RNA splicing. Evidence has shown that many genes have multiple forms of transcripts which are formed by different combinations of exons. With the progress in cDNA cloning, it has become clear that the majority of genes have multiple isoforms of transcripts. Some genes have been found to express more than a dozen of variants. Interestingly, a spliced variant is often found to express in a tissue-specific manner, and the proteins encoded by the transcriptional variants have diversified biological functions. The fact that a gene can express multiple transcripts increases the amount of genetic information carried by the gene. This phenomenon provides both opportunities and challenges for scientists who are trying to understand the biological functions of genes.

In recent years, the bioinformatic study has advanced tremendously along with the progress in the genome study, however, the efforts are put more on the sequence analysis and gene decoding fields. Desktop tools for analyzing alternatively spliced variants are extremely rare. With more cDNA sequences becoming available, molecular biologists now frequently encounter genes with multiple transcripts. To manage and understand the multiple transcripts requires sophisticated and user-friendly tools. GeneHarbor Inc. has devoted its resource to develop a tool for the purposes. We are now proudly to present ExonTracker™ 2.0. ExonTracker™ 2.0 is a windows system based software tool which enables users to analyze a typical genome Blast data much further. Some of the unique features in the package which, we believe, will enhance the ability of molecular biologists to study the exon-intron structure of a gene, make a graphic comparison of alternatively spliced variants, and evaluate the effects of an exon replacement on the protein functional domains. The innovative dot xon file ".xon" is extremely useful for designing PCR primers spanning exon junctions or within an exon.

We hope you will enjoy using ExonTracker™ 2.0 in your research, and we would also like to hear your opinions on the software and suggestions for future improvement.

Geneharbor, Inc.
January 31, 2004

Installation

System Requirements

The recommended system and configuration for ExonTracker™ 2.0:

Component	Minimum Requirement
Processor	Intel Pentium III or compatible 500 MHz or greater
RAM	256 MB or greater
Display	800 x 600 resolution, 256 color depth, small fonts setting, and 256 colors or greater.
Operating System	Windows 95, Windows 98, Windows Me, Windows NT 4.0, Windows 2000, or Windows XP
Drives	CD-ROM drive

CD-ROM Installation

If you have Autoplay turned on your computer will automatically run the CD-ROM interface, otherwise follow these directions:

1. Insert the CD-ROM into your CD-ROM drive.
2. From the Windows desktop, double-click the My Computer icon.
3. Double-click the CD-ROM icon.
4. Double-click the setup icon to start the installation interface.

Follow the instruction of each step during the installation process. ExonTracker 2.0 will be installed into a default folder assigned by this program. In case that you do not want to use the default folder, you may also install it to another location by choosing the custom installation method. All components are required for the program and should be kept in the designated folder all the time. After installing ExonTracker 2.0, the interface will automatically continue to install licenser device driver required to run ExonTracker 2.0. Follow the instructions to finish the process. An Icon for ExonTracker 2.0 will be placed on the desktop of your computer.

After the installation, attach the licenser key to your computer. Double-click on the ExonTracker™ 2.0 icon on the desktop to run ExonTracker™ 2.0.

Overview

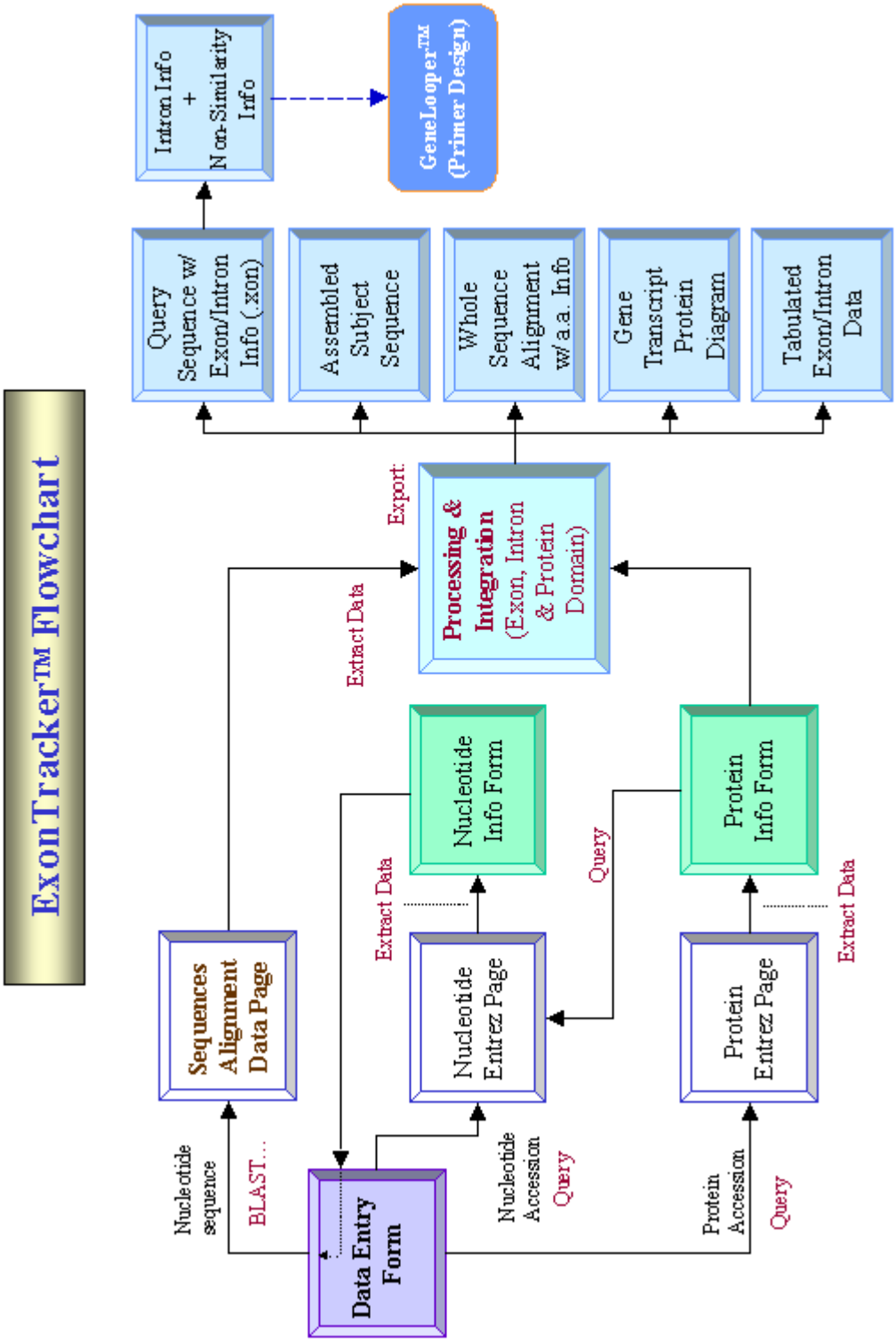
ExonTracker™ 2.0 is designed to assist researchers to understand how a transcript is assembled from RNA splicing. With the software, users can easily answer some basic questions, such as the number of exons to form a transcript, the distance between two adjacent exons (the length of intron), and the exons encoded a particular protein domain, and so on.

ExonTracker™ 2.0 fully relies on the resource and data in the public databases created by National Center for Biotechnology Information (NCBI). Particularly, it uses the returned BLAST data from a query (transcript) against a genomic database. It also uses the annotated information presented in the Entrez documents both for nucleotide and protein sequences. While the data from NCBI have already been providing tremendous help for researchers, ExonTracker™ 2.0, taking advantages of the processing ability and flexibility of a user's computer and user's intelligence, provides means for users to analyze sequence data even further based on their needs and get maximum out of a query sequence.

In order to complete the tasks, ExonTracker™ 2.0 retrieves three pieces of information related to a transcript sequence, including the nucleotide Entrez, the protein Entrez and the alignment data obtained by Blasting the transcript against a genomic database, and then processes and integrates the data to creates a dynamic and graphic rich model depicting the gene, the transcript and the protein. Using the model, one can easily identify the number of exons composing of the transcript and the lengths of the introns in the gene, and more importantly, the exon or exons corresponding to a particular protein domain. The most remarkable thing about the software is its ability to export data in multiple formats which are long sought by researchers. One example is the so-called dot xon file. It is a transcript sequence in fasta format mosaicked with the intron information. With this simple format, the exon/intron information of a transcript can be stored, transmitted and reproduced with a simple interpreter, along with its sequence information. It is extremely useful for designing primers to produce PCR products to cover multiple exons or to be within a single exon. The program also can precisely assemble the fragmented alignments provided by NCBI to create a whole sequence alignment following the order of the transcript and incorporating amino acid sequences translated from the transcript and the subject sequence, a useful feature for comparing a query sequence with the genomic sequence at the amino acid level.

Users who have done sequence BLAST and Entrez querying will have little difficulty to use ExonTracker™ 2.0. The specialized Web browser in the software has many shortcuts to some frequently used querying pages. There are convenient links in ExonTracker™ 2.0 for transferring data from one type to another. Users can perform data analysis online or offline using previously saved genomic BLAST data and Query nucleotide and/or protein Entrez documents.

To assist users to understand the logic and data flows in ExonTracker™ 2.0, we have created a Flowchart shown on Page 6 to give users an overview about the software. In the flowchart, the data sources and functional operations are presented and linked by arrows and lines. The detailed procedure for each utility can be found in the late chapters of the manual.


ExonTracker™ Flowchart
Data Entry Form
Nucleotide sequence
BLAST...
Sequences Alignment Data Page
Extract Data
Nucleotide Entrez Page
Nucleotide Accession
Query
Nucleotide Info Form
Extract Data
Query
Protein Entrez Page
Protein Accession
Query
Protein Info Form
Extract Data
Extract Data
Export
Processing & Integration (Exon, Intron & Protein Domain)
Query Sequence w/ Exon/Intron Info (.xon)
Assembled Subject Sequence
Whole Sequence Alignment w/ a.a. Info
Gene Transcript Protein Diagram
Tabulated Exon/Intron Data
Intron Info + Non-Similarity Info
GeneLooper™ (Primer Design)

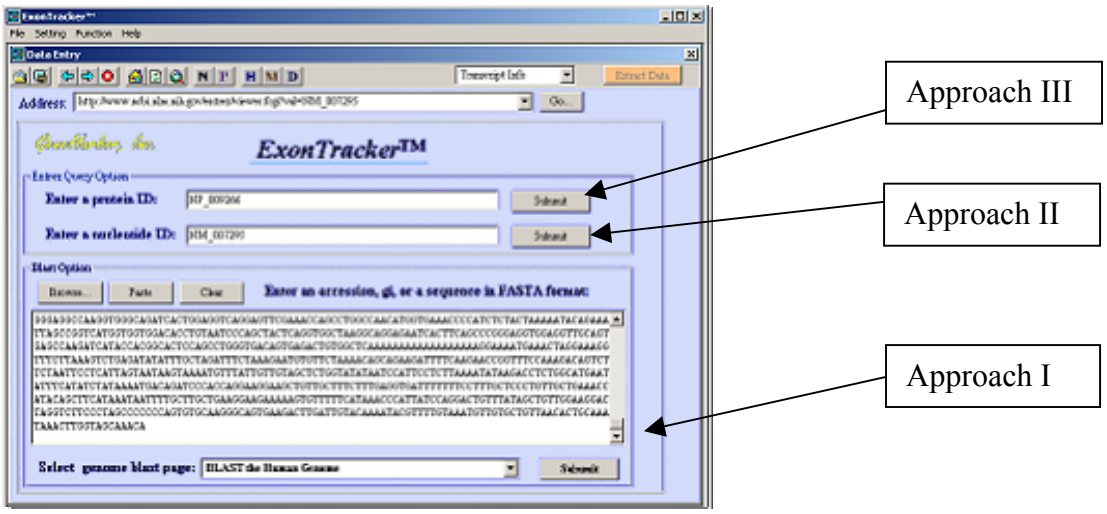
Protocol

A key step in establishing the relationship between a transcript and its genomic sequence using ExonTracker™ 2.0 is to upload the nucleotide sequence of a transcript into the program system. The sequence information serves at least two purposes. First the system uses it to detect the length of the transcript, its open reading frame (ORF), and subsequently uses the data to draw diagrams in various functionalities. Second the sequence or its ID is used as a query for genomic Blast. In order to perform the analysis, a user must have one of the following three things: 1) a nucleotide sequence in fasta format, 2) an accession number (ACCN) or a gene bank identification number (GI) of the transcript, or 3) the ACCN or GI for the protein that the transcript encodes. Based on the information you possess, select one of the three approaches described below to begin.

I. Preparation for launching the genomic BLAST

Launch ExonTracker™ 2.0 by double-clicking on the program icon on the Desktop of you computer. The DATA Entry form will appear (Fig. 1). Based on the initial information you have, choose one of the three approaches described below to begin.

Fig. 1. The **Data Entry** form with the three types of querying examples



Approach I: Beginning with a nucleotide sequence in fasta format:

1. Paste a nucleotide sequence to the text box from the clipboard or use the **Browse…** button to load a sequence file stored in your computer. It is now ready to Blast a genomic database.

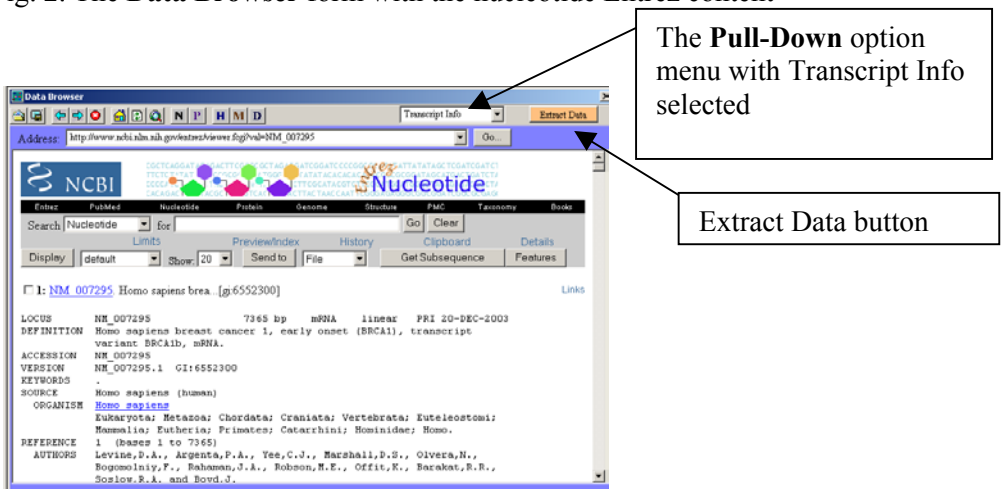
Approach II: Beginning with a nucleotide accession number:

1. Enter a nucleotide accession number in the text box titled "Enter a nucleotide ID:" in the DATA Entry form. Here we use the accession number for human breast cancer 1, early

onset (BRCA1), NM_007295, as an example to demonstrate the process, see Fig. 1. (its GI: 6552300 can also be used).

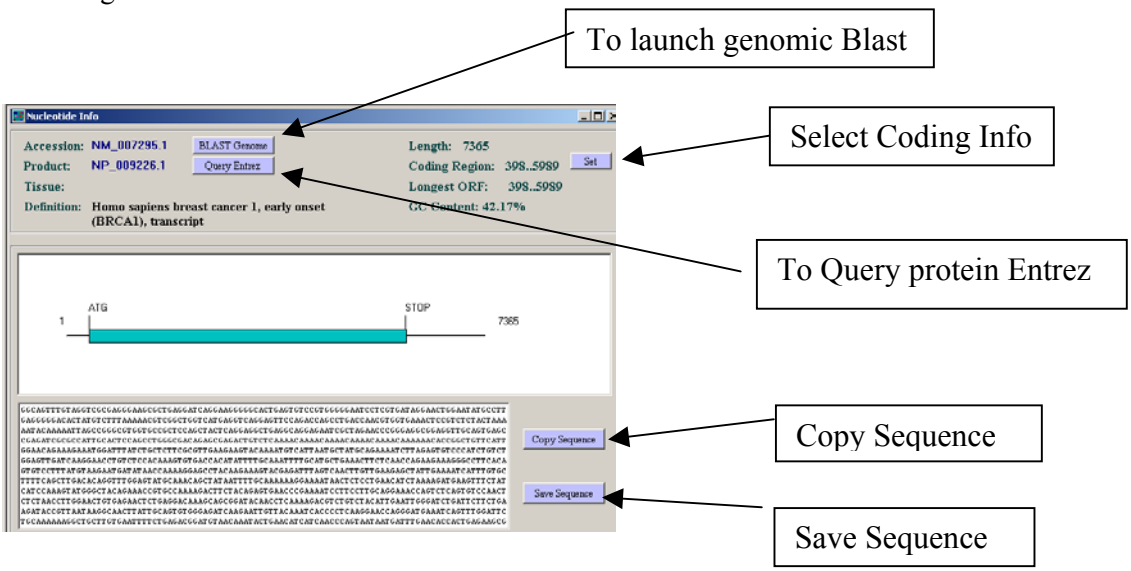
2. Click on the **Submit** button next to the text box. This will query the nucleotide Entrez database in NCBI and return the Entrez document of NM_007295 in the **Data Browser** form. This process may take several seconds to several minutes. Wait until the entire page is completely downloaded (Fig. 2).

Fig. 2. The **Data Browser** form with the nucleotide Entrez content



3. Select **Transcript Info** from the pull-down menu next to the **Extract Data** button. There are three options: **Transcript Info**, **Protein Info** and **Exon Info**. The program usually can detect the page contents, but it is recommended to make sure that the item selected corresponds to the content in the Data Browser form.

Fig. 3. The Nucleotide Info form with extracted data



4. Click on the **Extract Data** button to show the **Nucleotide Info** form (Fig. 3). There are three major parts in the form. The top part display several pieces of information about the transcript, including items directly read from the Entrez document (Access number, Protein ID, Description, Tissue, Length and the coding region) and two new items generated by the program, the longest open reading frame (ORF) and the GC-content.

The middle part of the form is a simple diagram indicating the length and the coding region of the transcript based on the information from the Entrez document. The lowest part is a text box containing the nucleotide sequence designed for copying and saving the nucleotide sequence.

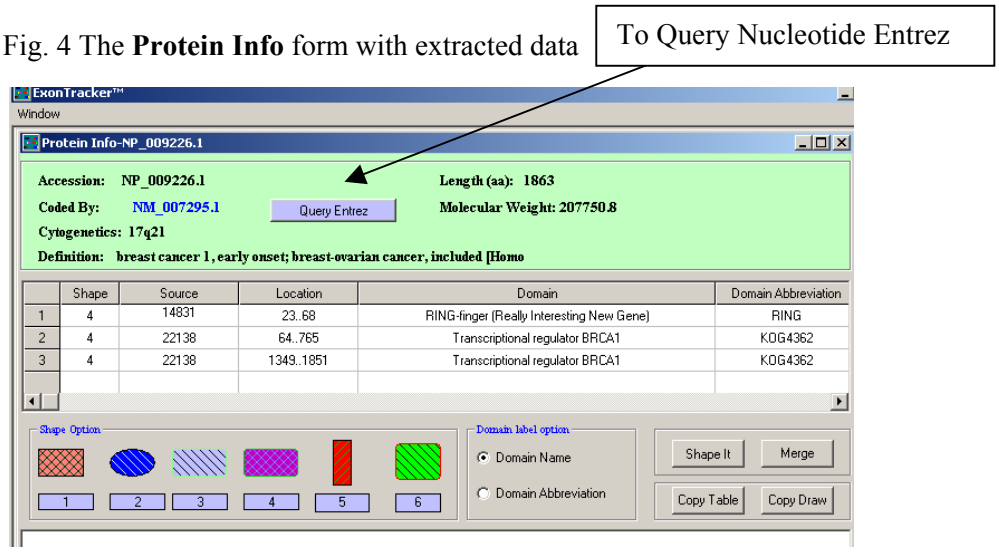
5. Click on the **To Genomic BLAST** button next to the nucleotide accession number. This action will input the accession number to the text box in the Data Entry form for genomic Blasting.

Note: You may wonder why just put the Accession number directly in the text box for genomic Blasting. The reason to go through the steps described above is to retrieve the physical sequence of the transcript from NCBI and stored it in the program system for late use.

Approach III: Begin with a protein Accession

1. Type the protein accession number in the text box titled as "**Enter a protein ID**". Here we use the accession number for human breast cancer 1, early onset (BRCA1), NP_009266, as an example. When its GI: 6552301 is used, the results will be the same (Fig. 1).
2. Click on the **Submit** button next to the text box. This will query the protein Entrez database in NCBI and return the Entrez document of NP_009226 in the **Data Browser** form (Figure not shown). Wait until the entire page is completely downloaded.
3. Click on the **Extract Data** button to show the **Protein Info** form (Fig. 4). In the top part of the form, there are several pieces of information about the protein, including items directly read from the Entrez form (Access number, Nucleotide Accession Number, Definition, Cytogentics, Length) and a new item generated by the program, the calculated molecular weight of the protein. Other contents of the form will be discussed in late chapters.

Fig. 4 The **Protein Info** form with extracted data



4. Double click on the Nucleotide accession number to query the nucleotide Entrez database in NCBI and retrieve the Entrez document of NM_007295. The **Data Browser** form with the nucleotide Entrez content will appear (Figure not shown).

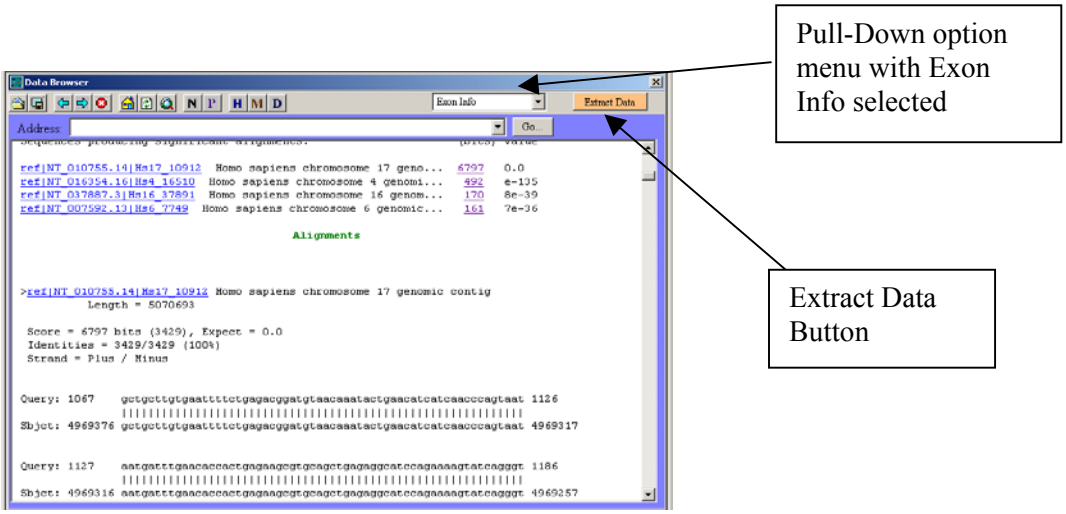
5. Follow the steps described in Approach II (begin from step 3) to complete the data extraction and submit the ID to the **Data Entry** form for Blasting.

II. Genomic BLAST

The operation of Genomic BLAST is the same as for the regular genomic Blast provided by BCBI. The **Data Entry** form serves just as a customized Web browser with many shortcuts to some frequently used genomic databases. The Genomic Blast links preset in the pull-down menu under the text box can be deleted, added and updated. Refer to Appendix A for additions or modifications.

1. Following the last step of all approaches described in previous chapter (Fig. 1), there should be a nucleotide sequence or an ID in the text box for BLAST. Select a desired genomic database from the pull-down menu under the text box.
2. Click on the **Submit** button to launch the BLAST. This action transfers the content in the text box of the **Data Entry** form to the input box of regular genomic Blast page.
3. Wait until the content appears in the input box for the standard genomic BLAST page provided by NCBI. Uncheck the **MegaBlast** option and set the Filter option to "None" because Blast with the two options sometimes result in the loss of short fragment alignments, consequently the loss of short exons.
4. Click on the **Begin Search** button to submit your blast request. This blast procedure is identical to the regular blasting procedures. Wait until the Blast data to return completely as the **Data Extract** button gradually becomes clear and the pull-down menu next to it displays "**Exon Info**". If it does not display the item, manually select "**Exon Info**" (Fig. 5).
5. Click on the Data Extract button to extract the Blast data and transfer it to the **Data Processing and Integration** form. It may take a few seconds to process the data and display the form if the file size is large.

Fig. 5. A sample of Blast Return in the **Data Browser** form



III. Data Processing and Integration

1. Wait until the **Data Processing and Integration** form appear (Fig. 6). The form has three parts. Its top part is a picture box for drawing diagrams. When the form initially appears, there is already a diagram depicting the query sequence with the length and coding region labels based on the transcript information. The middle part is designed to show sequence alignment dynamically in response to the movement of the mouse point within the picture box. The positions of the junctions of two adjacent aligning fragments and the locations of base pairs are also labeled in accordance with the alignment position. Click on the picture box once to stop the movement and click again to resume. The bottom part is a spreadsheet for storing and arranging the extracted data. The description of the column contents in the spreadsheet is shown in Table I. In addition to display the extracted data, the spreadsheet also serves as an operation panel for data processing.

Fig. 6. The **Data Processing and Integration form** with extracted sample data

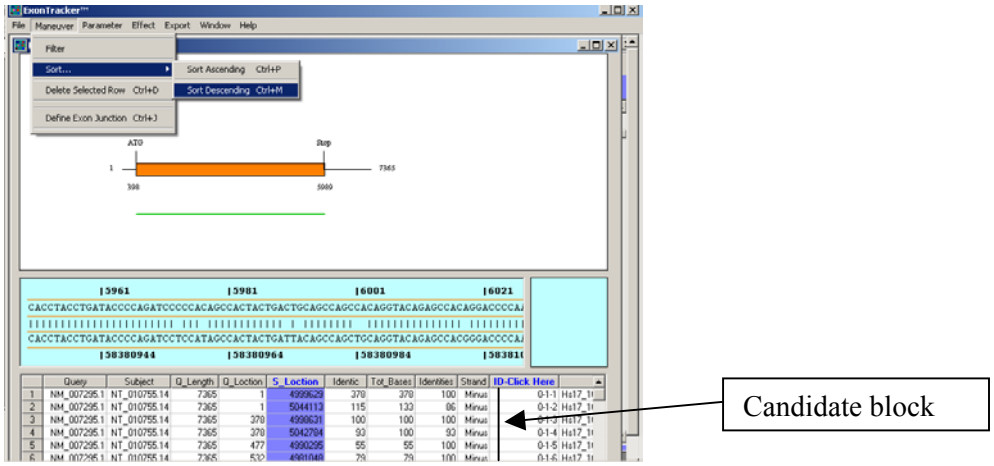


Table I. The descriptions of the column contents in the spreadsheet

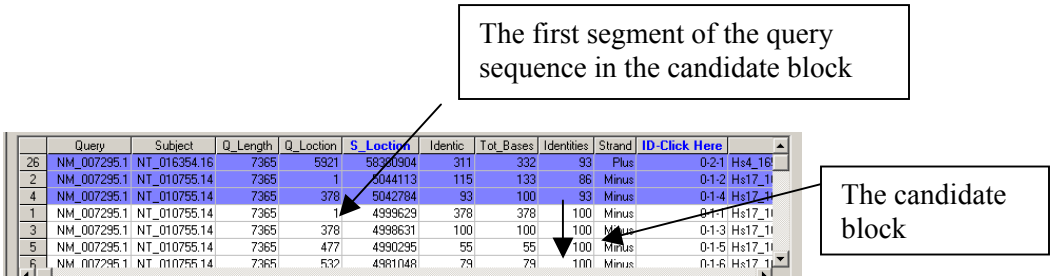
Column Name	Content	Note
Query	The accession number of a Query Sequence	Use Query_seq_1 if no accession #
Subject	The accession number of Subject Sequence	An Accession # is used for all aligned fragments found in the contig
Query_Length	The length of Query Sequence	
Q_Location	The first nucleotide position of query sequence in the segment	The column content is presorted ascending based on the position of a fragment in the query sequence and within a contig region
S_Location	The first nucleotide position of subject sequence in the segment	
Identical	Total identical bases between the two sequences	Data read from the Blast return
Tot-bases	The number of nucleotides of query sequence in the segment	Data read from the Blast return
Identities	Identical base divided by Total base	Data read from the Blast return
Strand	The strand of subject sequence	Data read from the Blast return
Id	Contig ID and Segment ID	Assigned by the program
Brief description	The description of each contig	Data read from the Blast return

2. Identify aligned segments corresponding to the real exons composing of the transcript. Follow the steps described below.

   Note: Identifying the aligned segments (Exons) and ordering them are the most critical operations in assembling the genomic sequence comprising the transcript. Based on our experience, the majority of transcripts are easy to be assembled, however some transcripts with multiple copies of homologous sequences in the same contig do require user's efforts. ExonTracker™ 2.0 provides many means to help users to deal with some very difficult situations.

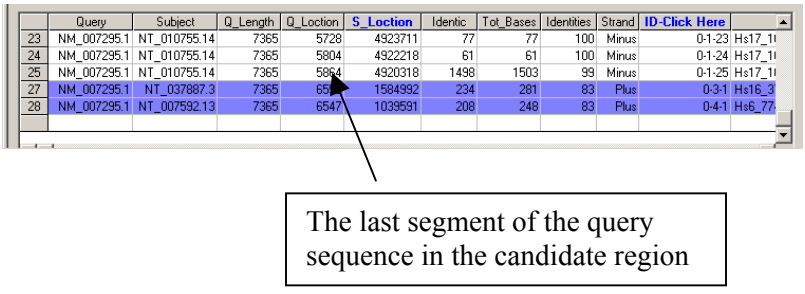
   a. Scroll through the spreadsheet and identify the block containing several segments with highest identities, usually being near 100%, as the candidate region in a contig. Remember the orientation of the subject strand: plus or minus.
   b. Click on the heading of **S_Location** (blue colored) to select the entire column.
   c. Open the **Sort…** menu under the **Maneuver** menu and select **Sort Ascending** if the orientation is plus, or **Sort Descending** if minus. The data in the spreadsheet will be sorted according to the **S_Location** column.
   d. Examine the **Q_Location** column and search for the smallest number (the first segment) within the candidate region, Fig 7. Highlight all rows above the row just identified. Because the selected rows are not true exon segments, delete them using the **Delete Row** function under the **Maneuver** menu.

Fig. 7. The spreadsheet with the upper non-exon segments selected after sorting the column **S_Location**



   e. Highlight all rows below the last segment (the last exon), judged by adding the number of **Q_location** to the number of **Tot_Base** and see if the sum roughly equals to the length of the query sequence (**Q_Length**) (Fig. 8). Delete the selected rows as described above because they are also non-exon segments.

Fig. 8. The spreadsheet with the lower non-exon segments selected after sorting the column **S_Location**



f. Click on the heading of **ID-Click Here** (blue colored) to select entire ID column. This action evokes the program to draw all exons (rectangles) and introns (lines) with labels of intron lengths above the existing transcript diagram. Each exon is pointed to the corresponding location of the query sequence by a pair of dotted lines. Step 'd' and 'e' remove all unrelated segments (non-exon matches) upstream of the first exon and downstream of the last exon, but not the unrelated matches inside the exon region. By examining the diagram, one can remove them manually if there is more than one genomic rectangles pointing to the same location of the transcript, indicating a redundant (non-exon segment) in the region. Delete the row to eliminate the non-exon segment. Click on the heading again to redraw the diagram. The next step can detect a redundant segment between two true exons automatically.

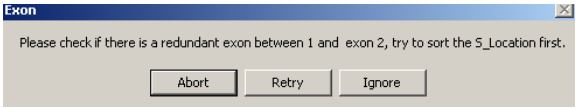
g. Select **Define Exon Junction** under **Maneuver** menu. By examining the overlapping sequences between two adjacent exons, the program predicts the splicing junctions based on the splicing rules, and then remove the overlapping base pairs. The program processes all junctions starting from the row representing the 5' of the query sequence.

Note: a) If there is no more redundant segment in the region, at the end of the process, the background color of the middle portion will make a change to indicate the data process is completed. The data in the spreadsheet and the location labels in the alignment display will be updated to reflect the segment data after the trimming.

b) If the program detects a gap (a **missing exon)**, it will give a warning message indicating the total number of the missing base pairs and its location, and then add "x" to fill the gap.

c) If the program detects a long stretch of overlapping base pairs, indicating there is a potential redundant exon in between, it will give a warning message to tell the location of the potential redundant piece and give users three options (Fig. 9). **Abort** will close the **Data Processing and Integration** form. **Retry** will return to the status before the **Define Exon Junction** was used. In this case, previously made deletions and sorting will remain effective. **Ignore** will continue the process despite the warning. Select **Retry** and see if there is indeed a redundant exon between the two segments detected by the program. If there is one, delete it using the **Delete** function mentioned above. Repeat the above process until there is no more redundant segment in the region. If you cannot find a redundant segment in the location detected by the program, it may be that the overlapping is naturally long and is treated as if there was a redundant segment by the program (false warning). Repeat **Define Exon Junction** and ignore the warning by clicking on **Ignore**. The process will proceed to pass this junction by removing the overlapping base pairs.

Fig. 9. Warning message box

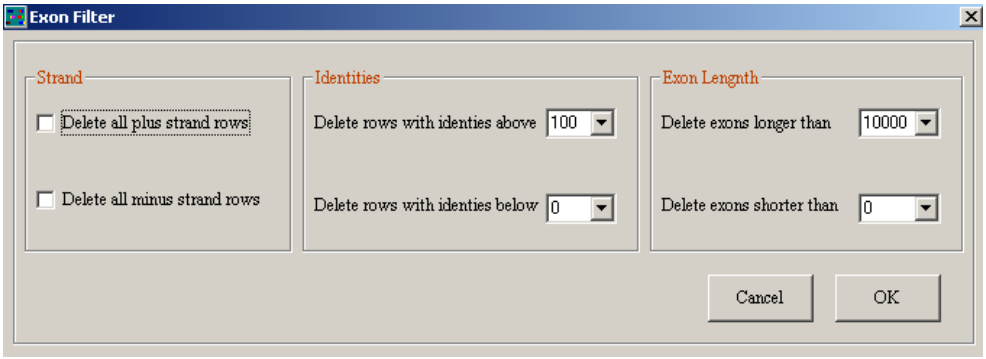


d) The **Define Exon Junction** utility completes two things, removing overlapping base pairs and predicting the exact splicing junction. The program tries to make the best judgment based on the sequence in the overlapping region. For the first goal, it can correctly remove the redundant base pairs in each junction. For the issue of predicting

splicing junctions, if the number of the overlapping base pairs is larger than one, the prediction accuracy is near 100%, when it equals to one pair, since there is not enough information for making a prediction based on the splicing rules, the program randomly removes one base pair. In this case, some prediction will be off by one base pair. Please make a notice on this issue. If the position of a splice junction is very critical to your analysis, use other means to define it.

d) With the help of human intelligence, it is possible to correctly process transcripts with many repeats in the genome using ExonTracker 2.0, which are difficult to be resolved by other means. The **Filter** function under the **Maneuver** menu is a very effective tool to deal with difficult transcripts (Fig. 10). The **Strand** option deletes either all segments with a plus orientation or the opposite. Since the exons for one transcript should have the same orientation, this function can remove all unwanted segments with the opposite orientation of the real exon segments. The **Identities** option removes those alignments with lower identities, which usually are not true exon segments of the transcript. The **Exon Length** option is useful for removing short repeats.

Fig. 10. **Filter** form. Check **Delete all plus strand rows** for removing all segments with a plus orientation, and check **Delete all minus strand rows** for removing all minus rows. To remove segments within a range of **Identities** values, use the **Identities** panel to set the range. To remove the segments within a range of sizes, use the **Exon Length** panel to set the range. Click on **OK** to accept the setting, and **Cancel** to quit.



IV. Addition of Protein Domain Information to the Diagram in Data Processing and Integration form

This section describes how to add protein domain information to the diagram created during Data Processing and Integration. If you do not want the information in your analysis, you can go directly to **Data Export** because the protein domain information is nonessential to other operations in the analysis. The **Protein Info** form has been introduced briefly in a previous

section. Follow the procedure described below to draw a protein diagram with its domain information.

1. Use the method described previously to retrieve a protein Entrez content to the **Data Browser** form. Check the Check box next to CDD and click on Display button to get annotated protein domain information (Fig. 11). Click on the **Extract Data** button with the **Protein Info** item in the poll-down menu selected to display the **Protein Info** form (Fig. 12).

Fig. 11. Method to get the pre-annotated domain information.

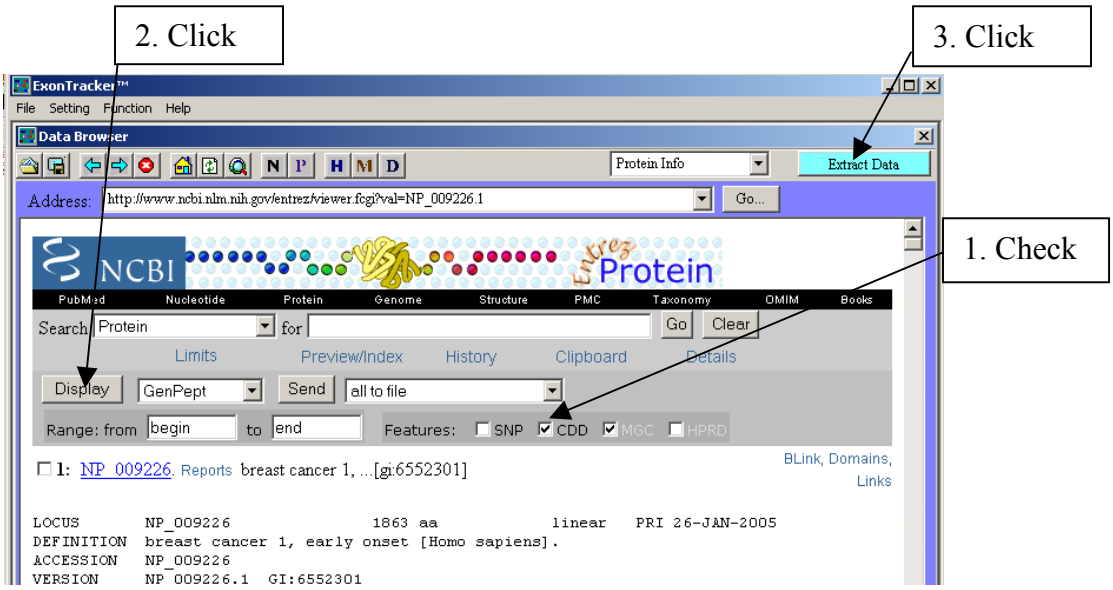
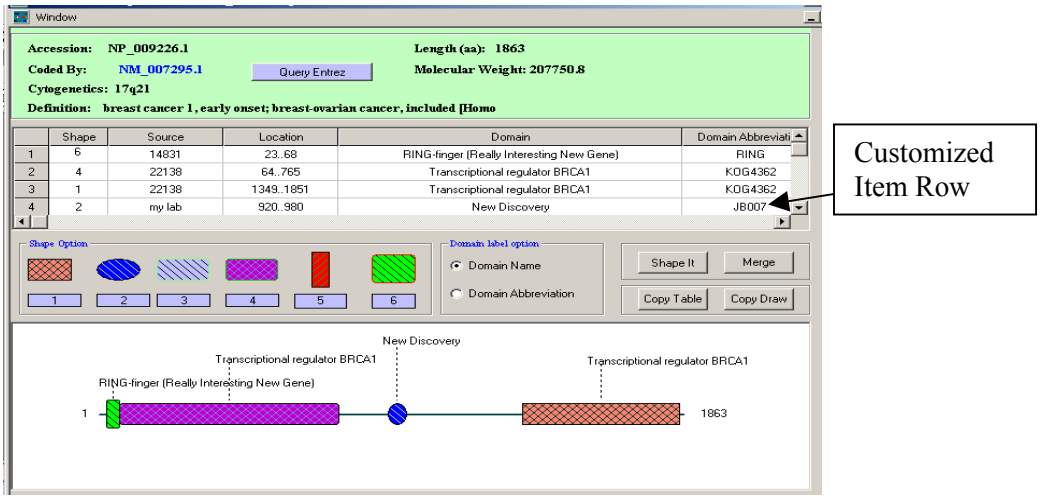


Fig 12. The **Protein Info** form with sample data. In the form, the general information about the protein read directly from the Entrez document is displayed in the top panel. It also includes the calculated molecular weight detected by the program. The annotated protein domain information in the Entrez is displayed in the spreadsheet. The program also detects hydrophobic regions in the protein using Kyte& Doolittle's method and displays a positive detection in the spreadsheet as one item.



2. The **Region** items in the protein Entrez are read and input to the spreadsheet. The data from one **Region** occupies one row and treated as one domain. Each one of the positive detections of hydrophobic regions is also treated as one domain. To add additional customized domains, input the data (Shape, Source, Location, Domain and Abbreviation) to the next row of the last domain row following the rules (See **Note** below).

Note: The format in the Location column is critical. The numerical number before ".." is the beginning point of the domain and the number after it is the ending point of the domain.

3. The number in the Shape column is the shape ID assigned by the program and can be changed manually to any one of the integer numbers available (1 to 6). To do so, just delete the old and type a new one. To hide a domain, delete the shape ID.
4. Use the **Domain Label Option** to select the content of Domain column or Domain Abbreviation as the domain label.
5. Click on the **Shape It** button to draw the domain diagram.

Note: The color and style of each preset shape can be customly modified. Use the follow method to make change on the preset shape.

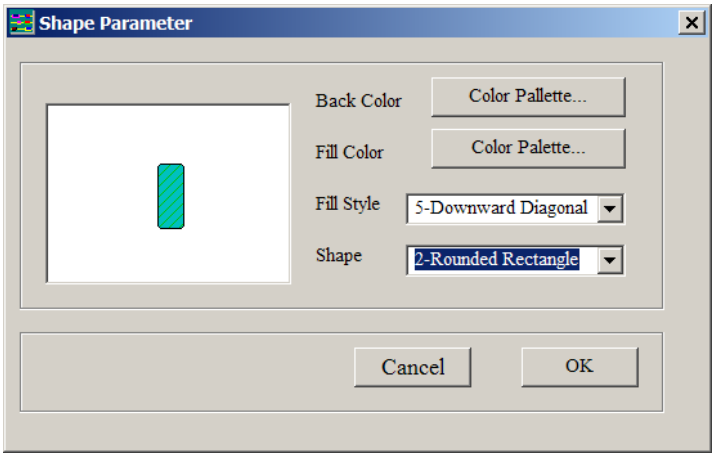
a) Select a shape by click on the shape ID under the shape. The shape is selected as indicated by the yellow-colored shape border.

b) Select the **Shape Color/Style** menu under the **Effect** menu. The Shape Parameter form will appear (Fig. 13).

c) Use the buttons and pull-down menus to adjust the parameters of the shape: back color, fill color, fill style and shape.

d) Click on the OK button to accept the change or Cancel to exit without any change.

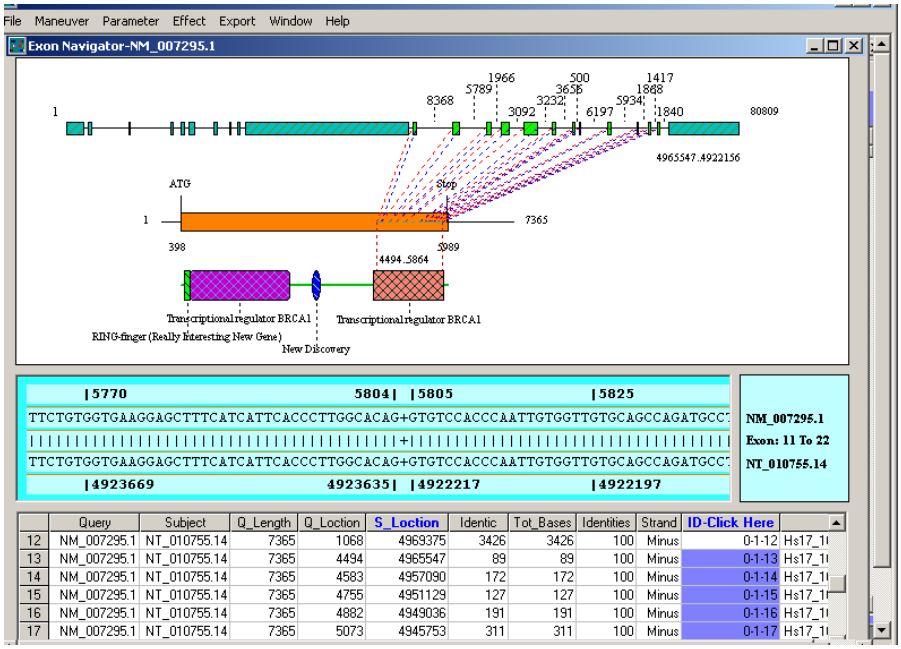
Fig. 13. The Shape Parameter form



6. To try different shapes for a domain, just change the shape ID by typing or double-clicking on the shape number after selected the number to be changed. Click on **Shape IT** to redraw the diagram.
7. To copy the diagram, click on the Copy Drawing button to the clipboard and then paste it to other picture editor such as MS PowerPoint and Adobe PhotoShop.

8. To copy the table content, highlight the rows to be copied and click on the Copy Table button.
9. Click on the **Merge** button to transfer the diagram to the **Data Processing and Integration** form.
10. The protein domain diagram will merge with the existing diagram in the picture area of the **Data Processing and Integration** form (Fig. 14). If the domain labels are stacked on each other, move the mouse point to the label, hold down the mouse left button and move the mouse to separate them. Click on the ID column to select row or rows corresponding to a domain. The dotted lines will connect exons with the corresponding region of the transcript and then with the protein domain. The panel in the right of the middle part will display the beginning and ending exons involved.
11. To hide the connection lines or the intron-labels, select the functions under the **Effect** menu correspondingly.

Fig. 14. Data Processing and Integration form with protein domain information

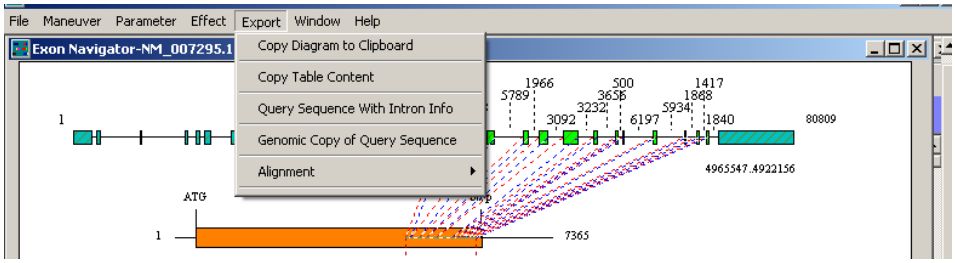


	Query	Subject	Q_Length	Q_Loction	S_Loction	Identic	Tot_Bases	Identities	Strand	ID-Click Here	
12	NM_007295.1	NT_010755.14	7365	1068	4969375	3426	3426	100	Minus	0-1-12	Hs17_1
13	NM_007295.1	NT_010755.14	7365	4494	4965547	89	89	100	Minus	0-1-13	Hs17_1
14	NM_007295.1	NT_010755.14	7365	4583	4957090	172	172	100	Minus	0-1-14	Hs17_1
15	NM_007295.1	NT_010755.14	7365	4755	4951129	127	127	100	Minus	0-1-15	Hs17_1
16	NM_007295.1	NT_010755.14	7365	4882	4949036	191	191	100	Minus	0-1-16	Hs17_1
17	NM_007295.1	NT_010755.14	7365	5073	4945753	311	311	100	Minus	0-1-17	Hs17_1

V. Data Export

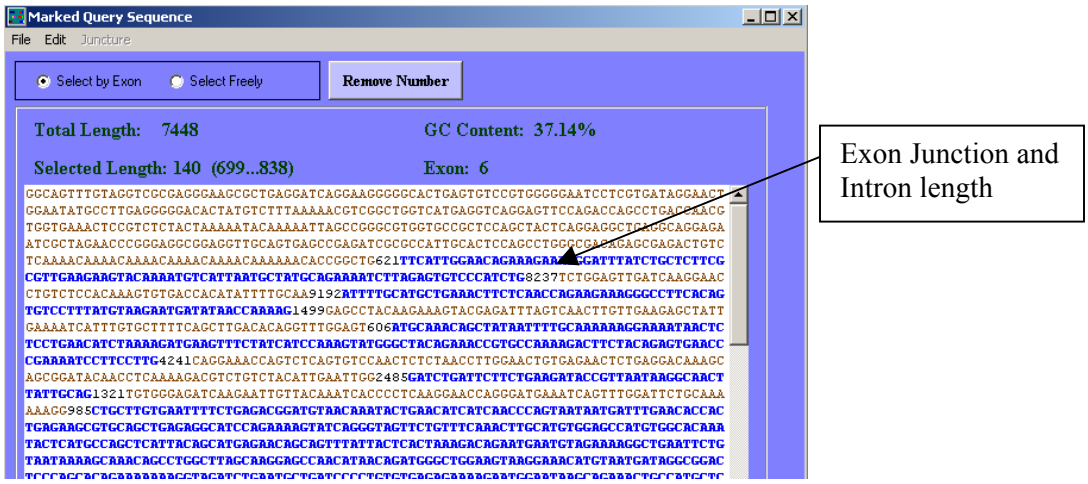
The Data extracted and generated by ExonTracker™ 2.0 can be exported in multiple formats which are designed for various research purposes and data presentations. The data export functions are organized under the **Export** menu in **Data Processing and Integration** (Fig. 15).

Fig. 15. The data export utilities under the Export menu in the Data Processing and Integration form



1. **Copy Diagram to Clipboard**: Select the menu under **Export**. The diagram in the picture box is copied to the clipboard and is now ready to be pasted to other picture environments provided by third party software packages including Microsoft PowerPoint and Adobe Photoshop.
2. **Copy Table Content**: Highlight desired rows or columns in the spreadsheet, and then select Copy **Table Content** menu under **Export.** The selected table contents are copied to the clipboard and can be pasted to Microsoft Excel sheet.
3. **Save The Spreadsheet**: Select **Save Data Table** under the **File** menu to save the spreadsheet as an Excel file if there is MS office installed in the computer.
4. **Query Sequence With Intron Info**: This function must be done after the **Define Exon Junction**. To use this exclusive data format created by GeneHarbor Inc., select the menu under **Export** to display the **Marked Query Sequence** (form Fig. 16).

Fig. 16. The **Marked Query Sequence** form with a sample "dot xon" file

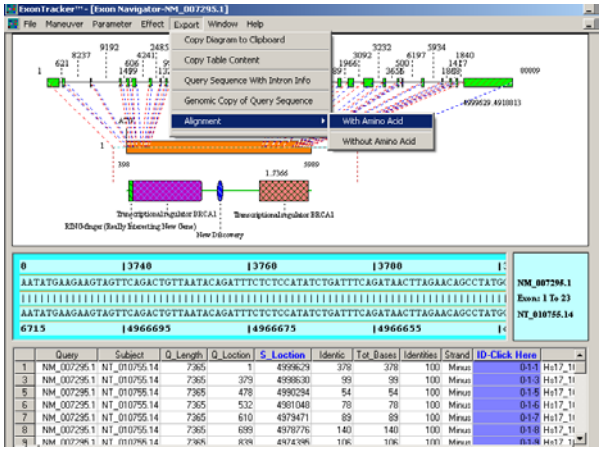


a) When the option of **Select by Exon** is selected, click on the sequence of the nucleotide will highlight the entire exon sequence, while when **Free Selection** is effective, selection will not be confined within an exon. To copy the select sequence, select the **Copy** menu under **Edit.** The copied sequence can be pasted to other text editor.

b) To save the sequence file, select the Save as to open the standard Save As window. The file can be saved as ".xon", "Doc", ".seq" or ".txt" file.
c) To print the file, select the **Print** function under **File** menu.
d) Click on the **Remove Number** button to remove all numbers inserted in the sequence.

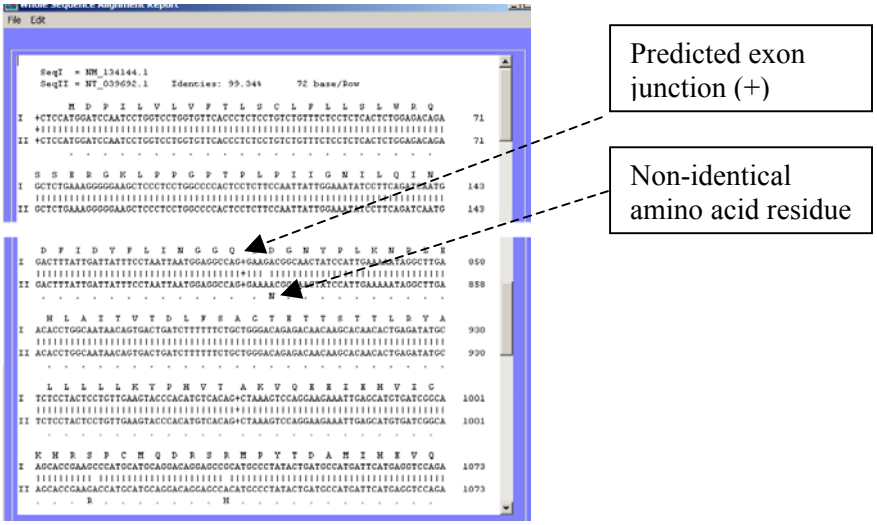
5. **Genomic Copy of Query Sequence**: The assembled genomic sequence matching to the transcript is exported by selecting the menu under **Export.** The sequence is displayed in the text editor similar to the previous one. Each exon region is colored and can be selected by exon or freely (Figure not shown). Similar to the previous file, it can be copied, saved and printed.
6. Export whole sequence **Alignment-With Amino Acid** sequences: This function can only be done after finishing **Define Exon Junction** and having all Exon rows selected (Fig. 17).

Fig. 17. The required status of **Data Processing and Integration** form just before exporting whole sequence alignment with amino acid labeled.



a) Select **Export-Alignment-With Amino Acid** under **Export** to display the whole sequence alignment in a text editor (Fig. 18).

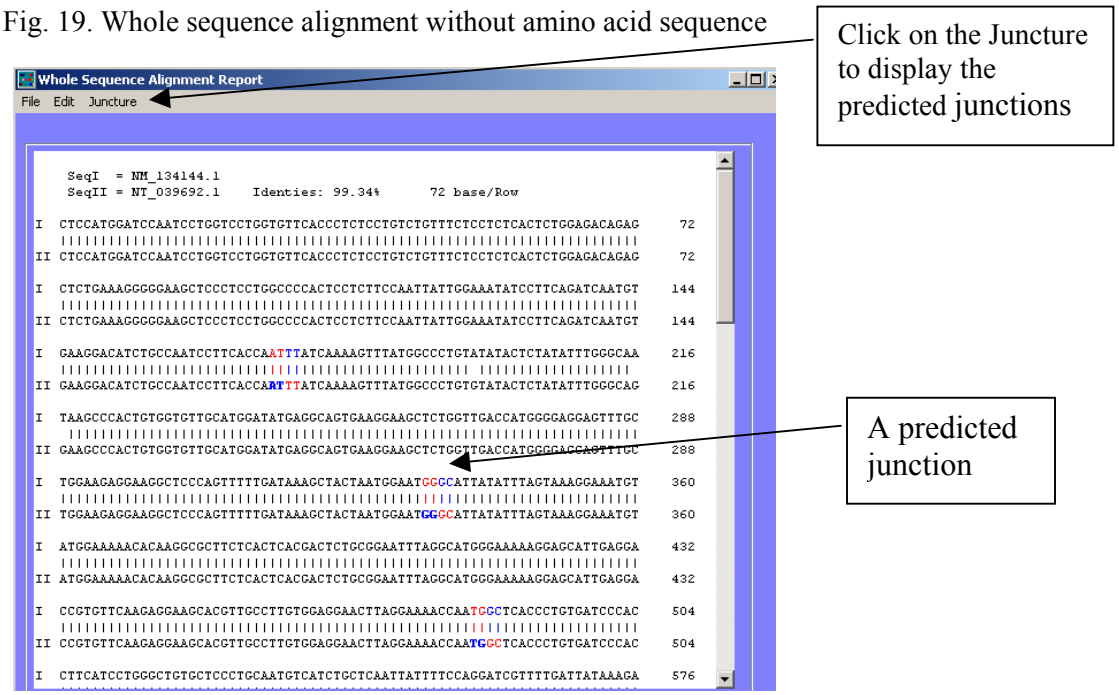
Fig. 18. Sample of exported whole sequence alignment with amino acid sequences. Click on the "|" rows to show the exon number.



b) The **select** and **copy** functions menus are under **Edit,** while **Save** and **Print** are under the **File** menu.

7. To export a whole sequence no labeled amino acid residues, select **Alignment-Without Amino Acid**. Note: unlike the previous function, it can be used to display the alignments of any selected number of exon segments (Fig. 19).

Fig. 19. Whole sequence alignment without amino acid sequence


Whole Sequence Alignment Report
File Edit Juncture
SeqI = NM_134144.1
SeqII = NT_039692.1 Identies: 99.34% 72 base/Row
Click on the Juncture to display the predicted junctions
A predicted junction

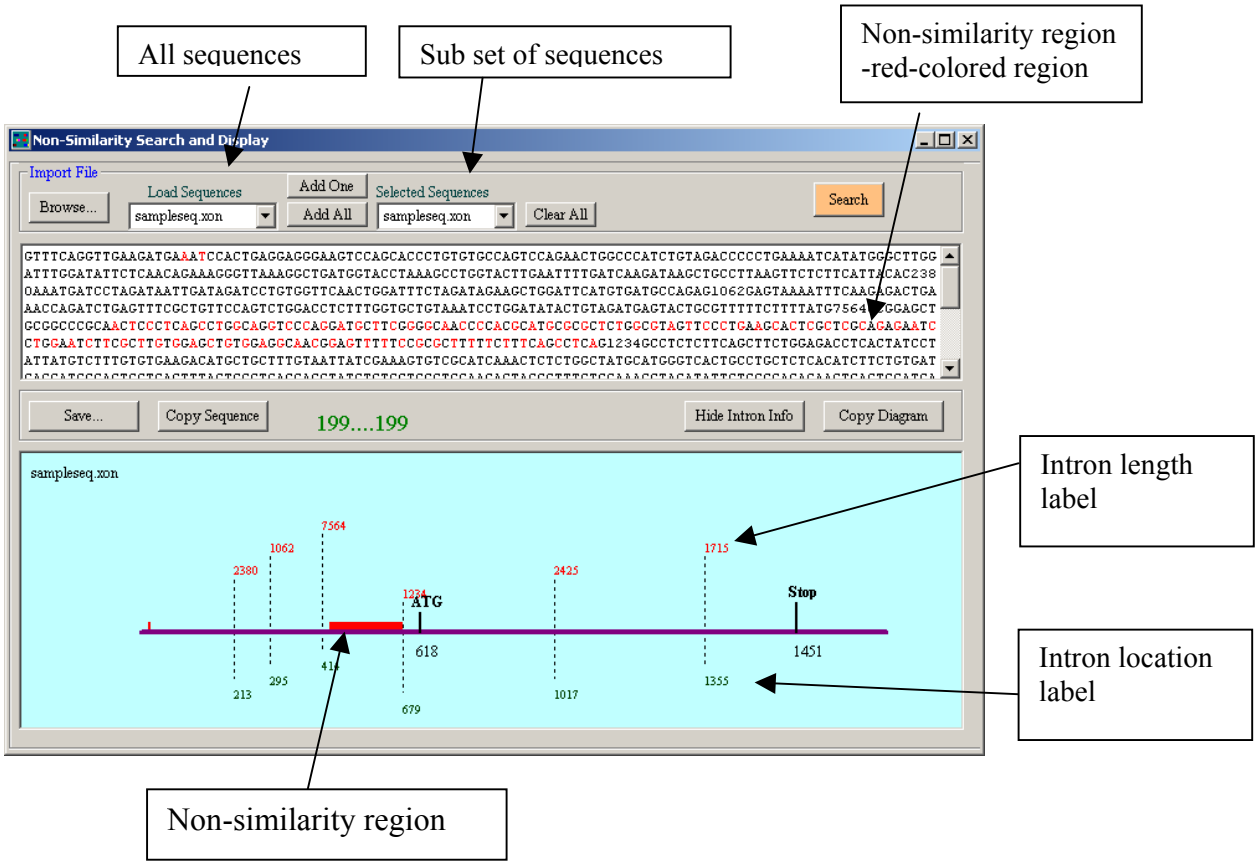
VI. Non-Similarity Search and Display

The Non-Similarity Search and Display utility is designed for users to search the non-similar regions of a sequence against a pool of related sequences. It is intended to identify unique exons in a transcript among a group of alternatively spliced variants, however, it also can be used to identify non-homologus regions of a sequence against the transcripts from a gene family or random sequences. This function has an practical use. As mentioned before, the dot xon file is a sequence format containing exon-intron information. With the **Non-Similarity Search function**, users are now able to create a sequence file with both exon-intron and non-similarity data. With this sequence file, it is possible to design primers to amplify DNA fragments not only spanning introns, but being unique to a splicing variant. The **Non-Similarity Search and Display** in ExonTracker 2.0 can process either ".seq" files created by other programs (no intron and exon information) or ".xon" files created using ExonTracker 2.0.

Procedure:

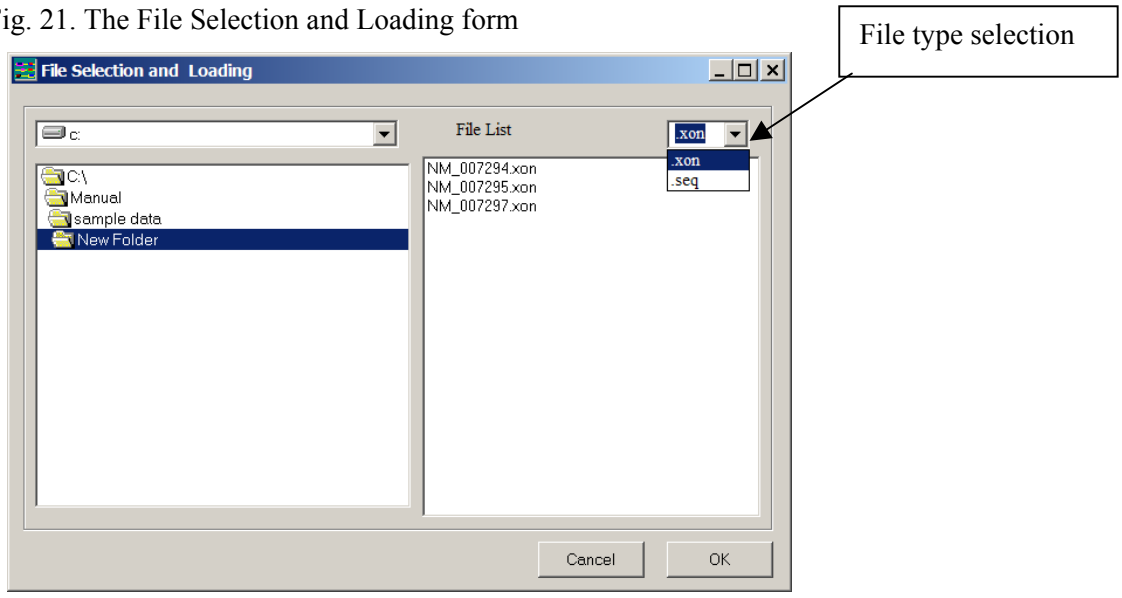
1. Create a folder and save all related sequence files in folder.
2. Start ExonTracker 2.0 program if it's not started and select **Non-Similarity Search and Display** under the **Function** menu of the ExonTracker 2.0 main form. The **Non-Similarity Search and Display** form will appear (Fig 20).

Fig. 20. The Non-Similarity Search and Display form with sample data



3. Click on the **Browser** button to show the **File Selection and Loading** form (Fig. 21), a special file open form for loading ".xon" files or ".seq" files.

Fig. 21. The File Selection and Loading form



4. Use the Drive browser to locate the folder containing the sequence files to be analyzed. The file names in the folder will appear in the file list box. Use the file type option menu above the file list box to select either ".seq" or ".xon" extensions.
5. Click on the **OK** button to load all files with the selected file extension in the folder to the system. The form will close and the file names will be input to the **Loaded Sequence** pull-down menu (Fig. 20). It contains file names loaded and can maximally hold 2,000 names.
6. All or a sub set of loaded sequences can be used in a search. Click on Add All to use all sequences for searching. To select a sub set of the sequences, use the Loaded Sequences pull-down menu to select the sequence and click on Add One button. The selected sequence name will appear in the Selected Sequences poll-down menu. One by one, add all desired names to the Selected Sequences pool. Use the pull-down menu under Selected Sequences to select a sequence you wish to analyze. The sequence will appear in the text box below. This sequence is used as the query and the other sequences who's names in the Selected Sequences will be used as subjects to be compared with.
7. Click on the **Search** button. The program starts to perform non-similarity search and may take from few seconds to several minutes depending on the number of sequences in the selected pool. When the process ends, the identified non-similarity regions of the sequence will be marked with red lines in the diagram. If the used sequences are dot xon files, the intron-lengths and their locations in the transcript sequence are labeled, thus the program creates a diagram showing the coding information, the non-similarity information and the exon-intron information (Fig. 20). The non-similarity regions of the sequence in the text box are also red-colored.
8. Click on the **Copy Diagram** button to copy the diagram to the clipboard, and then paste it to other picture editing environments such as PhotoShop and MS PowerPoint.
9. To save the sequence with intron and non-similarity information, click on the **Save Sequence** button to show the Save file window. Select the ".doc" as the file extension to save the sequence as a Microsoft Word document file. This sequence file can be read by the Primer Design in our previously released package, GeneLooper 2.0. With the sliding bars in the Primer Designer utility of GeneLooper 2.0, one can conveniently design primers to be specific to a unique region and to span introns. This is an extremely useful feature for studying the expressions of spliced variants.

VII. Primer Design - Take Exon Junctions Into Consideration

RT-PCR has been widely used to detect gene expression. A successful DNA amplification is partly depended upon the pair of primers used. Optimized primers can increase the yield of the amplified DNA and reduce the background caused by non-specific reactions. In addition to the general criteria, the locations of the pair of primers are also critical for producing reliable data. Specifically, designing primers to span exon junctions can eliminate false positive data due to amplifying contaminated genomic DNA in the mRNA used in the reaction. The primer design utility in ExonTracker is developed based on the data from thousand of PCR reactions and has been proven to be very reliable for designing optimal primers. Combining with the transcript sequence annotated with exon junctions and or our exclusive primer design layout, one can easily design optimal primers producing a fragment spanning exon junctions, thus obtaining unequivocal gene expression data.

Procedure:

Note: The primer design utility in ExonTracker 2.0 can processes either sequence files (.seq) or dot xon files (.xon). This function can be accessed through **Data Entry** under **Function** menu (Fig. 22A) or **Data Processing and Integration** under **Export** menu (Fig. 22B). The following procedure shows how to start primer design using the later after the completion of Define Exon Junctions.

Fig. 22A

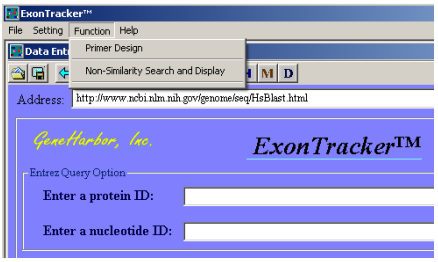
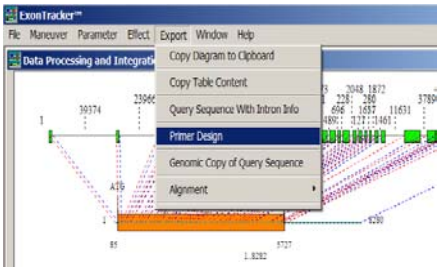


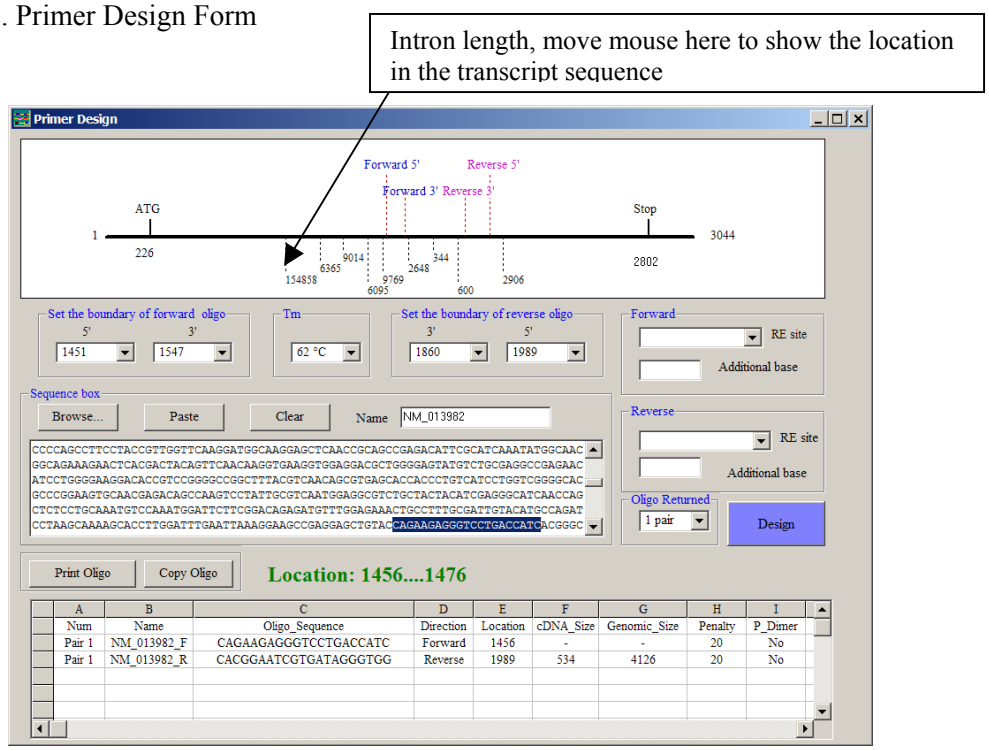
Fig. 22B



1. Upon the completion of Define Exon Junction (page), select Primer Design under Export in **Data Processing and Integration form**. The query sequence with exon junctions annotated is transferred to the Primer Design (Fig. 23).

Note: In the picture area of the form, there is a horizontal line depicting the length of the input sequence. The coding region of the sequence is marked by lines labeled with ATG and Stop. Each exon junction is indicated by a vertical line below the horizontal line, and the corresponding intron length is also labeled. On top of the horizontal line, four vertical lines (guidelines) are set in place for defining the regions to select primers. The lines can be moved by pointing the mouse to a line label, then holding down the left mouse button and moving it to a desired location. The position of each line can also be adjusted precisely by using the corresponding pull-down menus located just below the picture area.

Fig. 23. Primer Design Form



2. Define the boundaries of the forward primer and reverse primer. Use the two pairs of guidelines to set the boundaries. To design a primer pair to span exon junctions, move the four guidelines accordingly so that exon junctions are between the two 3' guidelines. One can directly set boundaries to a single base precision using the four pull-down menus.

3. Set a desired annealing temperature for the pair of primers using the Tm pull-down menu.

4. For a subcloning purpose (optional), you may add enzyme sites from the enzyme selection pull-down menu (**RE site**) for forward and reverse primers. These enzyme sites listed in the menu do not exist in the sequence between the 5' of the forward boundary and the 3' end of the reverse boundary, and they are dynamically updated following any change of positions of the two guidelines. Add a few bases to the 5' end of the restriction site to ensure a complete digestion of the PCR product with the selected enzyme.

5. Select the number of primers pairs to be designed on the **Oligo Returned** pull-down menu.

6. Click on the **Design** button to begin. It takes a few seconds for a primer pair to be displayed in the spreadsheet. The best pairs are always listed at the top of the table.

7. Clicking on the primer sequence in the table will highlight the primer sequence in the sequence box so that you can verify the primer sequence and examine the adjacent bases.

8. Highlight the rows you want to copy and then click on the **Copy Oligo** button. The copied rows can be pasted to other document environments such as MS Excel.

9. Click on the **Print Oligo** button to print the spreadsheet containing the primer information.

Notes:

1. A saved sequence file (".seq" or ".xon") can also be loaded directly by using the Browse button.
2. Any change in the sequence will trigger an update of the ORF information and the restriction enzymes list in the pull-down menu.
3. The program uses an arbitrary scoring system to evaluate the primers. The lower the penalty score, the better the primers.
4. There are sixteen pre-selected, frequently used enzymes for the 5'-addition.

BamH I/GGATCC	Sac I/GAGCTC
Bgl II/AGATCT	Sac II/CCGCGG
EcoR I/GAATTC	Sal I/GTCGAC
Hind III/AAGCTT	Sca I/AGTACT
Kpn I/GGTACC	Sma I/CCCGGG
Not I/GCGGCCGC	Spe I/ACTAGT
Pst I/CTGCAG	Xba I/TCTAGA
Pvu II/CAGCTG	Xho I/CTCGAG

Appendix A

Web Link Update

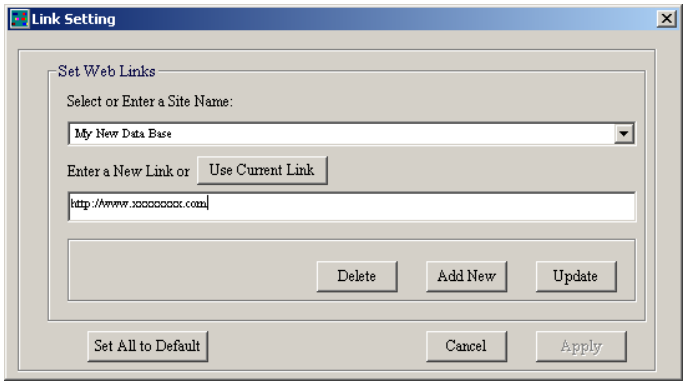
The URLs of the preset links in the package are frequently used by researchers and managed by NCBI. The collection may not be so broad to accommodate every user's needs, and they are subject to future changes by their administrators. This package includes the Web Link Update utility to give a user's ability to add new URLs or modify the URL of a preset link.

Procedure:

1. Start ExonTacker™ 2.0 by clicking on the program icon on the PC Desktop if the program is not open.
2. Select Web Link (URL) under the **Setting** menu to show the Link Setting form (Fig. 24).
3. To add a new link, paste or type the link name in the text box just under **Select or Enter a Site Name**, then type or paste the URL to the text box under **Enter a New Link**. You also can click on **Current Link** to input the current Web link showing on the Data Browser form. Click on **Add New** button, then **Apply** button. The new link is added to the links stored in the pull-down menu for genomic Blast.
4. To Update the URL of an existing link, select the name from the pull-down menu under **Select or Enter a Site Name** and then type or paste, or use the **Use Current Link** button to input the updated URL to the URL box. Then click on **Update**, then **Apply** button. The URL of the selected link is updated.
5. To delete an existing link, select the link name and click on the Delete button, then **Apply** button.
6. To change all links to the original setting provided by the package, click on **Set All to Default,** then **Apply** button.

Note: Any change made without clicking on the **Apply** button will not be effective. All changes made will be immediately effective after clicking **Apply** and remains so after restarted the computer.

Fig. 24. The **Set Web Link** form



Appendix B

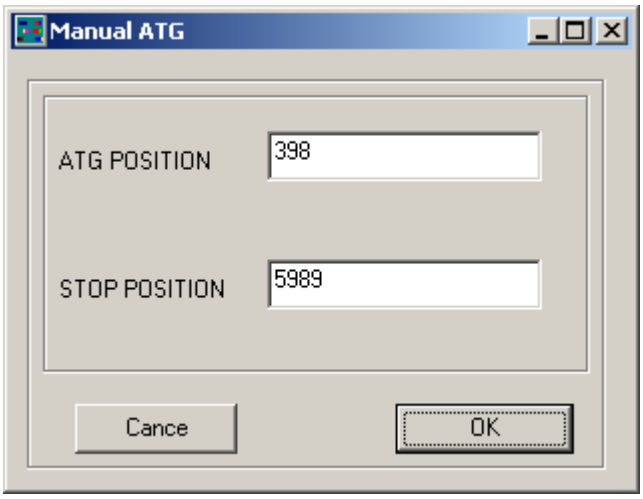
Manual Selection of A Coding Region

The coding region of transcript is the region encoding the protein, as defined by the initiation codon (ATG Position) and the termination codon (Stop Position). The coding information used in the analysis is very important in many diagrams. The coding region is detected as the longest open reading frame (ORF). It is also read from the Entrez document if it is available. If the coding region detected by the program defers from that read from the Entrez document, the program uses the late for the coding region. To use the longest ORF or other region, manually type the ATG and Stop positions.

Procedure:

1. Select **Set the Coding Region Manually** under the **Parameter** menu in Data Processing and Integration form. The form will appear (Fig. 25).
2. Manually type the numbers of ATG location and Stop location in the labeled text boxes. Then click on OK. The system will use the input coding information.

Fig. 25. Manual ATG form for inputting a desired coding region



Appendix C

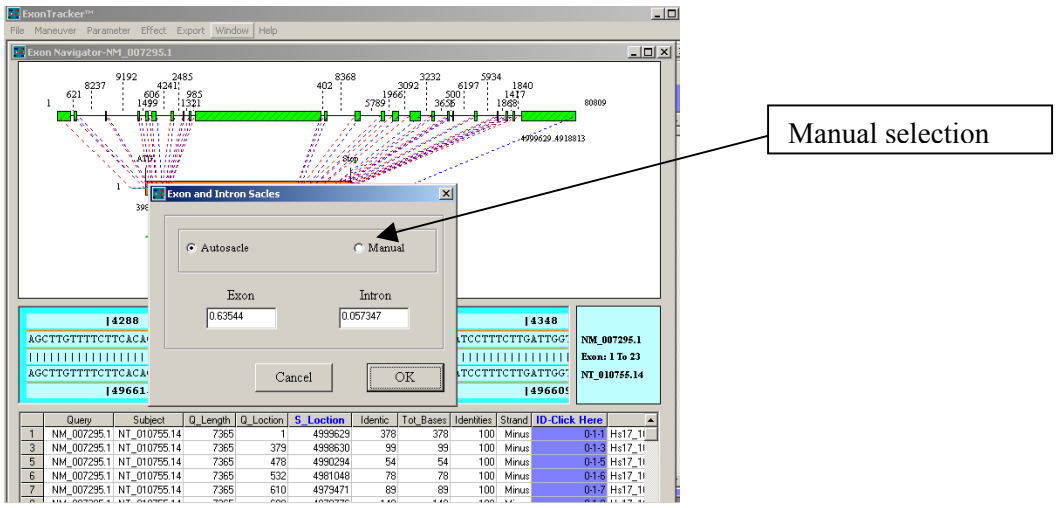
Set Exon-Intron Scales

The diagram in the Data Processing and Integration form is drawn using the scales calculated based on the length of transcript and the length of the genomic sequence involved. To reflect the relative sizes of an intron and exon, the program uses two different scales, one for the exons and one for the introns. The program automatically sets the scales to draw the diagram so that it fits well to the drawing area. In order to compare more than two transcripts graphically, it is better to draw the two in the same set of scales. The program includes a utility to give users option to set the scales manually.

Procedure:

1. Select Exon/Intron Scales under Setting menu of the Data Entry form or under Parameter menu of Data Processing and Integration to show the Exon/Intron Scales Form (Fig. 26).

Fig. 26. Setting the Exon/Intron scales manually



2. Click on the Manual option. Make a small change on the number in the exon scale and intron scale boxes.
3. Click on **Ok** to exit, Click on the heading of ID Click Here to redraw the diagram using the new scales. Try different scales until the dimension of the diagram is satisfied.
4. To draw two different diagrams in the same scale, Draw first diagram and open the Scale setting form and record the two numbers. Before draw the second, open the Scale form manually and input the scales detected from first transcript drawing, and then draw the second diagram. The two diagrams will have the same scales.